

К проблеме понимания несегментированного текста

(на материале метеорологических телеграмм)

Ирина Кононенко, Иван Попов

Российский НИИ Искусственного Интеллекта, Новосибирск

irina@mail.nsk.ru, popov@iis.nsk.su

Аннотация

В фокусе внимания данной работы находится проблема восстановления структуры несегментированного текста. Описывается процедура сегментации, которая осуществляет декомпозицию исходной лексической цепочки в последовательность тематически связанных фрагментов, в рамках которых возможна семантическая интерпретация. Определяется нарративная структура текста в терминах типа сценария (прогрессивный или рекуррентный) и структуры эпизода (параллельная или последовательная). Тематическая сегментация и анализ нарративной структуры текста позволяют установить семантические связи (сферы действия) локативных и темпоральных модификаторов.

Введение

Данная статья продолжает исследования по созданию модели понимания текстов в ограниченной предметной области на материале метеорологических телеграмм. Наш интерес к метеорологическим текстам, написанным в жанре телеграмм, имеет, по крайней мере, два источника. С одной стороны, существует практическая потребность в извлечении прогностической информации из таких сообщений и хранении ее в базе данных в структурированном виде с целью а) получения ответов на запросы пользователей, б) визуального представления в форме метеорологических карт. С другой стороны, широкое распространение спонтанно создаваемых текстов, характеризующихся отсутствием, либо недостатком сегментации, множеством сокращений, синтаксических и орфографических ошибок, ставит насущную задачу теоретических исследований и экспериментов с девиантными типами текстов.

Приведем типичный пример метеорологической телеграммы (М-текста):

погода томской на 19/08/98

переменная облачность утром местами туманы по южным мест небольшие кратковременн
дожди грозы по области ветер южный юго-западный 7-12 м/сек временами порывы до 16 м/с
т-ра ночью 8-13 днем 18-23 городе утром больш дожди ночью 10-12 днем 21-23

Корпус М-текстов демонстрирует характеристики подъязыка, особенности которого, как и знания о структуре предметной области, необходимо учитывать при создании модели понимания [1]. Первый этап эксперимента с М-текстами был представлен в [2], где описывались основные аспекты модели: структура предметной области, анализ специфики корпуса текстов, семантически-ориентированный подход к анализу (см. также [3,4]), представление лингвистической информации в виде иерархии классов и системы агентов, обеспечивающих лексические, пресемантические, тематические и семантические операции, связанные с обработкой основных структур входных текстов. За рамками рассмотрения в [2] остались вопросы сегментации текста и анализа обстоятельственных (локативных и темпоральных) объектов.

В фокусе внимания данной работы находится проблема восстановления структуры несегментированного текста. Сегментация текста предполагает декомпозицию исходной лексической цепочки в последовательность связных фрагментов, в рамках которых возможна семантическая интерпретация. Процесс сегментации М-текста описывается в разделе 2. В тесной связи с анализом структуры М-текста находится вопрос о выявлении семантических связей обстоятельственных модификаторов. Эта хорошо известная проблема (*modifier attachment*, см., например, [6]) приобретает при анализе несегментированного текста особую остроту: в отсутствие структурных показателей потенциальная сфера действия модификатора может быть очень широкой. Так, в приведенном выше примере М-текста темпоральный модификатор *утром* может быть потенциально семантически связан с целым рядом последующих прогностических оценок (*туманы, больш кратковременн дожди, грозы, ветер южный юго-западный 7-12 м/сек, порывы до 16 м/с*). Подход к определению реальной сферы действия обстоятельств описывается в разделе 3.

1. Базовые семантические компоненты модели

Модель понимания М-текстов интегрирует формальное описание структуры предметной области, которое лежит в основе выходных семантических представлений и в то же время позволяет организовать семантически-ориентированную процедуру анализа.

1.1. Систему *понятий предметной области* можно проиллюстрировать с помощью таблиц 1,2.

В структуре М-текста выделяются заголовок и главная часть. В заголовке указываются базовые локативный и темпоральный объекты - Территория и Дата. Главная часть представляет собой последовательность прогностических утверждений, структура которых описывается схемой «объект - признак - значение». Обстоятельственным (локативным и

темпоральным) объектам сопоставляются прогностические оценки в терминах Метеорологических Признаков, которые группируются в тексте вокруг Метеорологических Элементов (Осадки, Облачность, Ветер, ЯвленияПогоды, Температура), образуя тематически связанные фрагменты.

Таблица 1. Метеорологические понятия

МетЭлементы	МетПризнаки	МетЗначения
Осадки	ТипОсадков	Дождь, Мокрый снег
Облачность	НаличиеОблачности КоличествоОблачности	Есть, Нет Переменная, Облачно
ЯвленияПогоды	ТипЯвления	Туман, Гроза
Ветер	НаправлениеВетра ИзменениеВетра СкоростьВетра ПорывыВетра	Южный Юго-западный 7..12 м/с 16 м/с

Таблица 2. Локативные и темпоральные понятия

Объекты	ОбстПризнаки	ОбстЗначения
Лок	Территория Пункт ТеррОр ТеррЧасть ЛокКвант	Томская область Город Томск Юг, Юго-запад Горы, Верхнеленские районы Местами
Темп	Дата ДатЧасть ДатПодчасть Время ТемпКвант	19/08/98 Ночь, День Утро, Вторая половина 8.30 Временами

Соответствие между МетЭлементами и МетПризнаками представляется специальным прагматическим отношением.. Кроме того, ряд прагматических отношений представляют взаимосвязи обстоятельственных понятий: а) отношения «целое-часть», например, «ДатЧасть»- «ДатПодчасть»: {{«День» -- {«Утро», «Середина дня», «Вечер»}}, {«День» -- {«Первая половина», «Вторая половина»}},...}; б) отношение соответствия «Территория»-«Пункт»: {{«Томская область» — «Город Томск»}},...}.

1.2. Информационное содержание М-текста эксплицируется в виде *семантической сети*, узлы которой представляют понятия, а связи между узлами выражают *семантические отношения*, иерархия которых включает:

- *SA (МетЭлемент, МетПризнак)* — отношение Агрегации,
- *SD (Признак, Значение)* — отношение Доминации,
- *SP (ОбстПризнак, ОбстЗначение)* — отношение Целое-Часть,
- *SL (МетЭлемент, ЛокПризнак)* — Локативное отношение,
- *ST (МетЭлемент, ТемпПризнак)* — Темпоральное отношение.

Рисунок 1. Фрагмент семантической сети

На рис.1 представлена часть выходного семантического представления нашего примера, соответствующая фрагменту *по области ветер южный юго-западный 7-12 м/сек временами порывы до 16 м/с*. Заметим, что выходное представление М-текста содержит две объектные иерархии, которые строятся над базовыми объектами «Территория» и «Дата», а метеорологическая информация связана с ними посредством Локативного и Темпорального отношений. В [2] подробно описана семантически-ориентированная процедура анализа телеграмм, в результате которой получается метеорологическое дерево, изображенное в центре рис.1.

1.3. В основе семантически-ориентированного анализа лежит *механизм ориентаций*. Напомним, что *семантическая ориентация* лексемы указывает на множество Признаков, которые могут быть представлены лексемой в М-тексте. МетЭлементы рассматриваются как локальные темы текстовых фрагментов; сценарий описания каждой темы определяется множеством соответствующих МетПризнаков. *Тематическая ориентация* лексемы соотносит ее с множеством МетЭлементов, сценарии которых допускают данную лексему, т.е. имеют непустое пересечение с семантической ориентацией лексемы. Например, лексема *переменный* может представлять значение МетПризнаков «КоличествоОблачности» и «НаправлениеВетра», а тематически соответствует МетЭлементам «Облачность» и «Ветер». Соответствующая информация хранится в словарной статье лексемы в слотах СемОр и ТемОр.

В процессе семантически-ориентированного анализа текста осуществляются операции по лексической, пресемантической, тематической и семантической обработке текущей сети. Ниже рассматривается тематический анализ как метод сегментации М-текста и та часть семантического анализа, которая связана с обработкой обстоятельственных объектов.

2. Тематическая сегментация текста

Тематический анализ моделирует чтение лексической цепочки слева направо и осуществляет ее декомпозицию в последовательность тематически связанных фрагментов. По мере чтения порождаются узлы класса Тема, тематическая ориентация которых представляет тему (МетЭлемент) фрагмента, а составляющие фрагмент понятия (МетЭлементы, Признаки, Значения) связываются с Темой Тематическим отношением (ТН).

2.1. Продвижение по лексической цепочке.

В начале процесса порождается первый «активный» узел класса Тема, имеющий полностью неопределенную тематическую ориентацию (слот ТемОр содержит полный набор МетЭлементов). Текущая Тема связывается с очередным узлом цепочки при условии, что их тематические ориентации пересекаются (в том числе, если тема еще полностью не определена), что обеспечивает постепенное до уточнение темы.

Условием продвижения процесса является выполненность различных пресемантических операций, таких как сборка числовых интервалов и словокомплексов, локальные операции по снятию омонимии. Предварительная обработка обстоятельственных слов включает: а) сборку составных локативных имен вида «классификатор + идентификатор» (*Томская область, республика Тува*); б) установление кореферентности при прямой анафоре с помощью КоРеференциального отношения CR (*Томская область...по области*); в) восстановление смысла классификатора при косвенной анафоре (*Томская область...городе => 'Город Томск'*); г) предварительный анализ повторной референции с помощью Референциального отношения RR (*утром...утром, городе Томске...Томске*).

2.2. Тематический сдвиг

- создание нового «активного» узла класса Тема - происходит, если тематические ориентации очередного узла цепочки и текущей Темы не пересекаются: *по южным мест небольшие кратковременн дожди| грозы*.

Для М-текстов характерно введение новых тематических фрагментов обстоятельственными словами - описание МетЭлемента дается относительно конкретных обстоятельственных объектов: *погода томской на 19/08/98 переменная облачность | утром местами туманы*. В частности, последующие фрагменты дают дополнительное (уточняющее) описание того же МетЭлемента относительно новых обстоятельств: *по области ветер южный юго-западный 7-12 м/сек | временами порывы до 16 м/с; т-ра ночью 8-13| днем 18-23*. Это еще один случай тематического сдвига - новый узел Тема с неопределенной тематической ориентацией порождается при чтении очередного обстоятельственного слова (первого, если их несколько), причем новый узел получает ссылку на предыдущую Тему в слоте Супертема. По мере продвижения анализа и уточнения новой темы эта гипотеза о подтеме может быть отклонена, как во втором примере:

облачно [t1] | к утру прояснение [t2], где t1:Тема (Статус: «дезакт», ТемОр : {«Облачность»}, Супертема : t1),
t2:Тема (Статус: «акт», ТемОр : {«Облачность»}, Супертема : t1);

облачно [t1] | к утру местами туман [t2], где t1:Тема (Статус: «дезакт», ТемОр : {«Облачность»}, Супертема : t1),
t2:Тема (Статус: «акт», ТемОр : {«ЯвленияПогоды»}, Супертема : t2).

2.3. Контекстная верификация тематической сегментации

сопровождает процесс анализа. Тематический сдвиг означает, что прочитанный тематический фрагмент полностью «собран», но его тема может быть неточной. В таком случае необходима контекстная верификация, поскольку семантическому анализу подлежат

только полностью собранные тематические фрагменты с уточненной темой (Статус= «дезакт»).

Тематическая ориентация фрагмента уточняется его «активной» подтемой, если ее тема определена:

< утром местами туманы > | ночью небольшой [t1] | днем сильный снег [t2], где
t1:Тема (Статус: «собр», ТемОр : {«Облачность», «Осадки»}), t2:Тема (Статус: «дезакт»,
ТемОр : {«Осадки»}, Супертема : t1)
=> t1:Тема (Статус: «дезакт», ТемОр : {«Осадки»}).

Тематическая ориентация подтемы уточняется следующей «активной» подтемой:

днем без осадков [t] | ночью небольш [t1] | местами умеренные дожди [t2], где
t:Тема (Статус: «дезакт», ТемОр : {«Осадки»}),
t1:Тема (Статус: «собр», ТемОр : {«Облачность», «Осадки»}, Супертема : t), t2:Тема (Статус:
«дезакт», ТемОр : {«Осадки»}, Супертема : t)
=> t1:Тема (Статус: «дезакт», ТемОр : {«Осадки»}).

Тематическая ориентация подтемы уточняется предшествующей супертемой, в отсутствие следующей подтемы:

ветер южный 4-7 [t1] | ночью 10-12 [t2] | < небольшие дожди >, где
t1:Тема (Статус: «дезакт», ТемОр : {«Ветер»}),
t2:Тема (Статус: «собр», ТемОр : {«Ветер», «Температура»}, Супертема : t1)
=> t2:Тема (Статус: «дезакт», ТемОр : {«Ветер»}).

Заметим, что существуют ситуации тематической неоднозначности слов на границе тем (тематический стык): *сильный ветер сильн* [t1] | *метель* [t2] vs. *сильный ветер* [t1] | *сильн метель* [t2]. Левосторонний тематический анализ дает только первую - неверную - декомпозицию. Обычное решение с помощью локального синтаксического анализа (поиск хозяина прилагательного) неадекватно для текстов телеграмм, широко использующих сокращения, - более общим способом разрешения неоднозначности здесь являются прагматические факторы (порядок слов) и семантическая верификация.

2.4. Определение потенциальных сфер действия обстоятельственных слов

является частью процесса сегментации. Любому узлу класса Тема сопоставлены четыре списка ссылок на локативные и темпоральные Значения, которые потенциально или реально являются объектами прогностических оценок соответствующих тематических фрагментов: ПотСДЛок и ПотСДТемп, СДЛок и СДТемп. Списки ПотСД фиксируют границы, в рамках которых возможно распространение семантических связей ОбстЗначений. Левая граница для ОбстЗначения v1 соответствует содержащему v1 тематическому фрагменту. Очевидно, что v1 включается одновременно в списки ПотСД и СД «активного» тематического узла t1 этого фрагмента. Поскольку сфера действия v1 может распространяться вправо (t1 открывает потенциальную сферу действия v1), то v1 включается в списки ПотСД всех следующих фрагментов до t2, содержащего прагматически оппозитивное ОбстЗначение v2 (e.g. *днем* [v1]... *ночью* [v2], *по южным* [v1] *районам*... *по области* [v2]). Распространение потенциальной сферы действия v1 осуществляется при порождении новых тематических узлов:

городе [v1] утром небольш дождь [t1] | ночью 10-12 [t2], где
t2.Номер = t1.Номер+1,
t1:Тема (Статус: «дезакт», СДЛок Ê {v1}, ПотСДЛок Ê {v1}),
t2:Тема (Статус: «акт», ПотСДЛок : t1.ПотСДЛок).

по южным [v1] мест небольшие кратковременн дожди [t1] | грозы [t] | по области [v2] ветер южный юго-западный 7-12 м/сек [t2], где
 $t1.Номер < t.Номер < t2.Номер$,
 $t1:Тема (Статус: «дезакт», СДЛок \hat{E} \{v1\}, ПотСДЛок \hat{E} \{v1\})$,
 $t:Тема (Статус: «дезакт», ПотСДЛок \hat{E} \{v1\})$, $t2:Тема (Статус: «акт»,$
 $СДЛок \hat{E} \{v2\})$, $ПотСДЛок \hat{E} \{v1, v2\}$
 $\Rightarrow t2:Тема (ПотСДЛок : t2.ПотСДЛок - \{v1\})$.

Проверка реальности сфер действия ОбстЗначений, включенных в списки ПотСД, осуществляется на этапе семантического анализа.

3. Семантический анализ обстоятельственных объектов

В результате тематической сегментации текст представляется как последовательность тематических фрагментов с размеченными потенциальными семантическими связями эксплицитно упомянутых в нем обстоятельственных объектов. Верификация потенциальных сфер действия обстоятельств базируется на анализе нарративной структуры текста.

3.1. Нарративная структура текста

определяется типом сценария и структурой эпизода. Выделяется два типа *сценариев*: а) *прогрессивное* изложение - характеризуется последовательным введением и описанием различных тем (МетЭлементов); б) *рекуррентное* изложение - характеризуется возвратами к ранее введенным темам с целью их дополнительного описания. Приведенный пример М-текста иллюстрирует рекуррентный тип сценария: описание темы «Температура» прерывается возвратом к теме «Осадков». Легко представить себе изложение той же информации посредством сценария прогрессивного типа:

погода томской на 19/08/98 переменная облачность утром местами туманы по южным мест небольшие кратковременн дожди грозы городе утром небольш дождь по области ветер южный юго-западный 7-12 м/сек временами порывы до 16 м/с т-ра ночью 8-13 днем 18-23 городе ночью 10-12 днем 21-23

Эпизод представляет собой часть текста, посвященную описанию одной темы. Структура эпизода отражает способ изложения темы эпизода. В простом случае тема описана одним тематическим фрагментом. Если же тема описывается в рамках последовательности вида «супертема+подтемы», то можно говорить об одном из двух типов эпизодических структур. *Последовательная* структура предполагает, что супертема вводит описание темы относительно некоторого ОбстЗначения $v1$, вслед за которым в подтеме дается уточняющее описание относительно ОбстЗначения $v2$ (причем $v2$ не оппозитивно $v1$), например: *по области [v1] ветер южный юго-западный 7-12 м/сек [t1] | временами [v2] порывы до 16 м/с [t2]; днем[v1] без существенных осадков [t1] | второй [v2] половине неб дождь[t2]*. *Параллельная* структура характеризует эпизод, в котором тема описывается с двух сторон - описание относительно ОбстЗначения $v1$ дополняется описанием относительно прагматически оппозитивного ОбстЗначения $v2$, е.g. *ночью[v1] т-ра 8-13 [t1] | днем [v2] 18-23 [t2]*. Существуют комбинированные эпизодические структуры, в которых последовательное описание вложено в параллельное, е.g. *ночью[v1] т-ра 8-13 [t1] | городе [v2] 10-12 [t2] || днем [v3] 18-23 [t3] | городе [v4] 21-23 [t4]*.

3.2. Верификация потенциальных сфер действия ОбстЗначений

осуществляется в рамках семантического анализа. Прежде всего, проверяются эпизоды комбинированной параллельно-последовательной структуры. Реальная сфера действия ОбстЗначения v , находящегося во вложенной позиции, не должна распространяться на вторую часть параллельной структуры:

т-ра ночью [v1] 8-13 [t1] | городе [v] 10-12 [t2] | | [v2] днем [v1] 18-23 [t3], где
t1.Номер < t2.Номер < t3.Номер, v1.СемОр = v2.СемОр, t1.Супертема=t3.Супертема,
t1:Тема (СДТемп $\hat{E}$ {v1}),
t2:Тема (СДЛок $\hat{E}$ {v}, ПотСД Темп $\hat{E}$ {v1}),
t3: Тема (СДТемп $\hat{E}$ {v2}),
ПотСДЛок $\hat{E}$ {v}
=> t3: Тема (ПотСДЛок : t3.ПотСДЛок - {v}).

При проверке эпизодов последовательной структуры, сфера действия ОбстЗначений из предшествующих фрагментов распространяется на последующие:

т-ра ночью [v] 8-13 [t1] | городе 10-12 [t2], где
t1. Номер < t2. Номер, t1.Супертема= t2.Супертема,
t1:Тема (СДТемп $\hat{E}$ {v}),
t2:Тема (ПотСДТемп $\hat{E}$ {v}) => t2:Тема (СДТемп : t2.СДТемп $\hat{E}$ {v})

Несколько операций связано с проверкой рекуррентных нарративных структур. Обнаруживаются последовательности тематических фрагментов, темы которых повторяются далее, а также последовательности фрагментов, повторно описывающих ранее введенные темы. Сфера действия ОбстЗначения v распространяется из первого фрагмента такой последовательности на последующие, например:

т-ра ночью 8-13 [t1] | ...| городе [v] утром небольш дождь [t2] | | [v2] ночью 10-12 [t3], где
t1. Номер < t2. Номер < t3. Номер, t1.ТемОр = t3.ТемОр, t2:Тема (Супертема : t2,
СДЛок $\hat{E}$ {v}), t3: Тема (ПотСДЛок $\hat{E}$ {v})
=> t3: Тема (СДЛок : t3.СДЛок $\hat{E}$ {v}).

В результате операций верификации обнаруживаются и восстанавливаются имплицитные связи обстоятельственных объектов с последующими фрагментами текста. В данный момент эти связи представлены СД-списками, ассоциированными с тематическими узлами. Остается решить две задачи. Во-первых, СД-списки должны быть конвертированы в две объектные, локативную и темпоральную, иерархии. Во-вторых, эти иерархии должны быть связаны соответствующими отношениями с корнем каждого метеорологического дерева.

3.3. Создание объектной иерархии

рассмотрим на примере локативной иерархии для нашего М-текста. В нем вводятся 5 различных локативных объектов, которые включены в СДЛок - списки тематических узлов:

погода томской [v1] на 19/08/98
переменная облачность [t1] | утром местами [v2] туманы [t2] | по южным [v3] мест [v4]
небольшие кратковременн дожди [t3] | грозы [t4] | по области [v1] ветер южный юго-западный
7-12 м/сек [t5] | временами порывы до 16 м/с [t6] | т-ра ночью 8-13 [t7] | днем 18-23 [t8] |
городе [v5] утром небольш дождь [t9] | ночью 10-12 [t10] | днем 21-23 [t11]

t1: Тема (СДЛок :{v1}, ТемОр :{«Облачность»}); t2: Тема (СДЛок : {v2, v1}, ТемОр :
{«ЯвленияПогоды»}); t3: Тема (СДЛок : {v4, v3, v1}, ТемОр :{«Осадки»}); t4- t8: Тема (СДЛок :
{v1}); t9- t11: Тема (СДЛок :{ v5, v1}).

Для каждого ЛокЗначения v порождается ЛокПризнак f (если его нет в соответствующем тематическом фрагменте) и связывается с v отношением Доминации $SD(f, v)$. Далее, на множествах ЛокЗначений каждого СДЛок-списка устанавливаются отношения Целое-Часть более общих объектов к более конкретным и включаются в семантическую сеть (если они в ней еще отсутствуют):

$t2.СДЛок : \{v2, v1\} \Rightarrow SP(f1, f2);$
 $t3.СДЛок : \{v4, v3, v1\} \Rightarrow SP(f1, f3); SP(f3, f4);$
 $t9.СДЛок : \{v5, v1\} \Rightarrow SP(f1, f5).$

Наконец, всякий МетЭлемент el связывается Локативным отношением с наиболее конкретным объектом из СДЛок-списка соответствующего тематического узла:

$ТН(t1, el1), el1.Смысл: «Облачность» \Rightarrow SL(el1, f1);$
 $ТН(t2, el2), el2.Смысл: «ЯвленияПогоды» \Rightarrow SL(el2, f2);$
 $ТН(t3, el3), el3.Смысл: «Осадки» \Rightarrow SL(el3, f4);$
 $ТН(t4, el4), el4.Смысл: «ЯвленияПогоды» \Rightarrow SL(el4, f1);$
 $ТН(t5, el5), el5.Смысл: «Ветер» \Rightarrow SL(el5, f1);$
 $ТН(t6, el6), el6.Смысл: «Ветер» \Rightarrow SL(el6, f1);$
 $ТН(t7, el7), el7.Смысл: «Температура» \Rightarrow SL(el7, f1);$
 $ТН(t8, el8), el8.Смысл: «Температура» \Rightarrow SL(el8, f1);$
 $ТН(t9, el9), el9.Смысл: «Осадки» \Rightarrow SL(el9, f5);$
 $ТН(t10, el10), el10.Смысл: «Температура» \Rightarrow SL(el10, f5);$
 $ТН(t11, el11), el11.Смысл: «Температура» \Rightarrow SL(el11, f5).$

Остается прокомментировать еще два вопроса обработки обстоятельств. Первый касается отношения Целое-Часть на множестве ЛокЗначений. Доминация по этому отношению не известна заранее для Значений Признаков «ТеррОр» и «ТеррЧасть» и должна определяться из структуры текста. Информации о порядке следования Значений и их принадлежности к одному/разным тематическим фрагментам оказывается достаточно для определения доминирующего Значения:

на севере $[v1]$ верхнененских $[v2]$ районов... $[t]$, где
 $t:Тема(СДЛок \hat{E} \{v2, v1\});$
в верхнененских $[v2]$ районах... $[t1] \mid$ на севере $[v1]$... $[t]$, где
 $t:Тема(СДЛок \hat{E} \{v2, v1\}), t1.Супертема = t.Супертема,$
 $v1.СемОр = «ТеррОр», v2.СемОр = «ТеррЧасть»$
 $\Rightarrow v2(Дом:true), SP(f2, f1).$

Второй вопрос касается установления кореферентности обстоятельственных слов при повторях. В результате пресемантической обработки неанафорические повторы связаны с предшествующими употреблениями специальным референциальным отношением: $RR(v2, v1)$. Подтверждение тождества референции влечет замену Значения $v2$ Значением $v1$ в содержащих $v2$ СД-списках, например:

на севере $[v1]$ области $[v]$... $[t1] \mid \dots \mid$ на севере $[v2]$ <в верхнененских $[v3]$ районах>... $[t2]$, где
 $t1:Тема(СДЛок \hat{E} \{v, v1\}), v1.СемОр = «ТеррОр», v.СемОр = «Территория»,$
 $t2:Тема(СДЛок \hat{E} \{v, v2\}), RR(v2, v1),$
 $(v1.Дом=true \dot{\cup} (t1.СДЛок \dot{\cup} \{v3\})=\{ \}),$
 $(v2.Дом=true \dot{\cup} (t2.СДЛок \dot{\cup} \{v3\})=\{ \}), v3.СемОр = «ТеррЧасть»$
 $\Rightarrow t2:Тема(СДЛок : t2.СДЛок - \{v2\} \hat{E} \{v1\}).$

Симметричная проверка связана с повторной референцией к Значениям Признака «ТеррЧасть». Что касается Значений Признака «ЛокКвант», то условием тождества объектов является совпадение СДЛок-списков: $t1.СДЛок - \{v1\} = t2.СДЛок - \{v2\}.$

Аналогичная процедура создает иерархию темпоральных объектов. Заметим, что обе иерархии могут подвергаться дополнительной постобработке. Например, темпоральная иерархия для нашего примера должна локально модифицироваться с учетом прагматической информации «ДатЧасть» - »ДатПодчасть».

Заключение

Опыт анализа и извлечения информации из текстов метеорологических телеграмм показывает, что предложенный метод тематической сегментации дает, как правило, вполне удовлетворительные результаты. Развитие метода должно обеспечить разрешение неоднозначности слов на границе тем, анализ сочинения при отсутствии структурных показателей (союзов, знаков препинания). Кроме того, планируется модификация описанной процедуры в сторону локальной многовариантности для анализа лексических девиаций (при ошибках, сокращениях).

Дальнейшая разработка описанного подхода требует расширения экспериментальной базы - привлечение близких по жанру текстов из других предметных областей позволит не только апробировать предложенные тематические и нарративные механизмы на новом материале, но и расширить номенклатуру анализируемых нарративных структур.

Литература

Hein S. Definite Nps and Background Knowledge in Medical Text. In: Computers and Artificial Intelligence; Vol.8, N6, 1989, pp.547-580.

Загорулько Ю., Кононенко И., Попов И. Экспериментальная система понимания метеорологических телеграмм. // Труды международного семинара Диалог'99 по компьютерной лингвистике и ее приложениям, т.2 - Таруса, 1999 - с.57-66.

Нариньяни А.С. Автоматическое понимание текста - новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. - Москва, 1997, с.203-208.

Kononenko I., Sharoff S. Understanding Short Texts with Integration of Knowledge Representation Methods. In: Perspectives of Sysem Informatics, Lecture Notes in Computer Science; Vol. 1181, Springer, 1996, pp. 111-121.

Franz A. Automatic Ambiguity Resolution in Natural language Processing. An Empirical Approach. Lecture Notes in Artificial Intelligence; Vol. 1171, Springer, 199, 155 p.